

Deep generative-contrastive networks for facial expression recognition

Youngsung Kim[†], ByungIn Yoo^{‡,†}, Youngjun Kwak[†], Changkyu Choi[†], and Junmo Kim[‡]

[†]Samsung Advanced Institute of Technology (SAIT), [‡]KAIST

yo.s.ung.kim@samsung.com, byungin.yoo@kaist.ac.kr, yjk.kwak@samsung.com, changkyu.choi@samsung.com,
junmo.kim@ee.kaist.ac.kr

Abstract

As the expressive depth of an emotional face differs with individuals, expressions, or situations, recognizing an expression using a single facial image at a moment is difficult. One of the approaches to alleviate this difficulty is using a video-based method that utilizes multiple frames to extract temporal information between facial expression images. In this paper, we attempt to utilize a generative image that is estimated based on a given single image. Then, we propose to utilize a contrastive representation that explains an expression difference for discriminative purposes. The contrastive representation is calculated at the embedding layer of a deep network by comparing a single given image with a reference sample generated by a deep encoder-decoder network. Consequently, we deploy deep neural networks that embed a combination of a generative model, a contrastive model, and a discriminative model. In our proposed networks, we attempt to disentangle a facial expressive factor in two steps including learning of a reference generator network and learning of a contrastive encoder network. We conducted extensive experiments on three publicly available face expression databases (CK+, MMI, and Oulu-CASIA) that have been widely adopted in the recent literatures. The proposed method outperforms the known state-of-the-art methods in terms of the recognition accuracy.

1. Introduction

Facial expressions are a primary modality to understand the emotional status of an individual. The expression provides a useful contextual clue for social communication [11]. However, individuals do not always clearly reveal their facial expressions. When an individual reveals an ambiguous facial expression, a human may have an experience to compare his/her expression with other expressions observed in past in order to extract their facial expression differences. The related evidence is found in the literature of brain sciences. According to [4, 5, 11], an individual

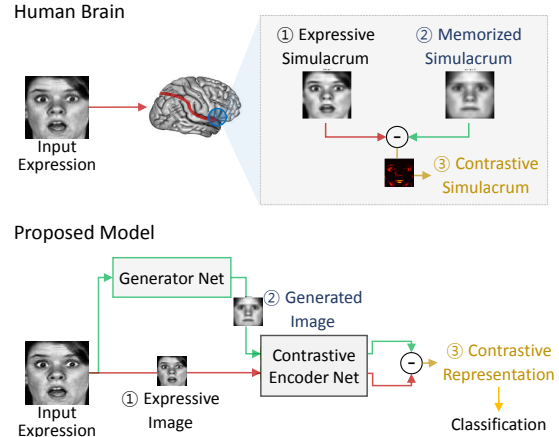


Figure 1: Overview of our proposed architecture. A similar procedure with the proposed architecture might be observed in a human brain. A given expressive face is compared with the memorized facial shape in the brain. In the proposed networks, a feature of the given expressive face is compared with that of a reference image which is estimated by a generator network.

can discern various facial expressions by recalling the memorized face shapes of a shown person. The neural pathways for detecting changeable aspects of faces (e.g., eye movements and emotional expressions) and for memorizing the unique face shape are separately distributed [4, 11]. These two processes are interacted in the core system of the brain [5, 11].

We attempt to utilize a reference face image that indicates the memorized unique face in the brain to discriminate a facial expression input in a deep neural network framework (see Figure 1). We assume that an expression factor can be extracted from the *contrastive characteristics* between the given image and the reference image. The reference image for an individual identity, however, is not always available in the wild. We start from the assumption that there is a generative artificial neural network that can be used to infer a reference image from the given facial ex-

pressions. If a single image is given, the reference image is generated using the generative (encoder-decoder) networks. The next required process is to model the *contrastive characteristics* mentioned above using deep neural networks.

One of main concerns is to find out how to extract or encode the contrastive features between the reference image and the given expression image. In our proposed networks, two representation models are included, for 1) disentangling of expressions and 2) explaining of contrastive representation with a supervised setting. In general, deep networks disentangle multiple variation factors of an input image. Several unknown or unintended factors are revealed in the networks and a useful factor is selected by a proper objective. In this paper, we attempt to disentangle directly the intended factor: a facial expressive factor. Hence, disentangling of expression is conducted in two steps. First, through learning a generator network that estimates a reference expression image, expressive factors can be eliminated. This estimated reference image is used to measure a expressive representation by comparing with the original expression image in the feature space. In a later part, disentangling is assisted by contrastive metric learning and a supervised reconstruction.

From the approach in the literature, gradual changes of facial expressions are utilized to extract temporal information along the multiple frames. This multiple images (video) based model has abundant information of the expression transition, which can be used for the recognition. In this paper, we focus on exploring a representation from a pair of a generated frame and a given frame. Our proposed framework could be easily extended to utilize multiple frames as well.

In this paper, we attempt to answer to a few questions quantitatively and qualitatively: 1) Is a generated reference image useful for the discriminative task? 2) How generative networks are controlled by contrastive metric learning for a discriminative purpose? 3) How does facial generation affect expression recognition?

The main contributions of this paper are as follows:

- We combine encoder-decoder networks and convolutional neural networks into a unified network that simultaneously learns to generate, compare, and classify samples on a dataset.
- We show that the contrastive representation trained with contrastive metric learning and a supervised reconstruction is useful to achieve a better discriminative performance for a facial expression recognition task.
- We show that the proposed method outperforms the state-of-the-art methods including the multiple images based approach in terms of facial expression recognition accuracy even when a single image is utilized in a test phase.

2. Related works

Facial expression recognition has been studied over decades. Several different approaches exist that are based on local feature extraction, facial action units (FAUs), temporal information, and convolutional neural networks. The local feature-based methods such as the Gabor filter, LBP, HOG, and BoW are the most common and widely studied to extract good visual features [1, 15, 27]. In the FAU based methods [16, 17], FAUs are detected and analyzed to classify an expression. This is mainly based on the facial action coding system (FACS) proposed by Paul Ekman [7]. Temporal information-based methods [19] utilize multiple images. These methods, however, achieve *limited* recognition accuracy performance because the designed features lost some information. To overcome the insufficient representations of the hand-crafted features, deep learning based methods have been recently adopted. An ensemble of two deep networks models that handle temporal information including appearance and geometric features has been proposed [13]. A simple convolutional neural network has been used to analyze the FAU in the learned filter of the networks [14]. To obtain discriminative spatiotemporal representation, facial action parts detection is performed using 3D-CNN [18]. However, it shows limited performance when compared to the state-of-the-art methods. This is because those CNN-based methods still could not show a good enough representation of a facial expression.

Another deep learning framework has been proposed to take advantage of the discriminative and generative models for realizing a better generalization performance. Traditionally in generative networks such as the autoencoder, a popular approach is that the entire stack of encoders is finetuned using pre-trained autoencoders in a layer-wise manner for discriminative purposes. Recently, a generative model was simultaneously learned with a discriminative model. In generative adversarial networks (GANs) [8], the generative model is learned against an adversary and a discriminative model that learns to determine whether a sample is from the model distribution or data distribution. The stacked what-where auto-encoders (SWWAEs) [28] integrate discriminative and generative learning pathways and provide a unified approach to supervised, semi-supervised, and unsupervised learning. In this paper, we deploy a generative model with discriminative learning as well. We are mainly focusing on investigating a contrastive representation of a facial expression that is optimized with appropriate objectives.

3. Contrastive facial representation learning

Consider an input image matrix \mathbf{X} and a reference image matrix \mathbf{X}_r that are elements of a given set of image matrices $\mathcal{I} = \{\mathbf{X}_i \in \mathbb{R}^{h \times w} \forall i\}$. The corresponding expression label is denoted by $y \in \mathbb{R}$ and $y_r \in \mathbb{R}$ respectively. In a real

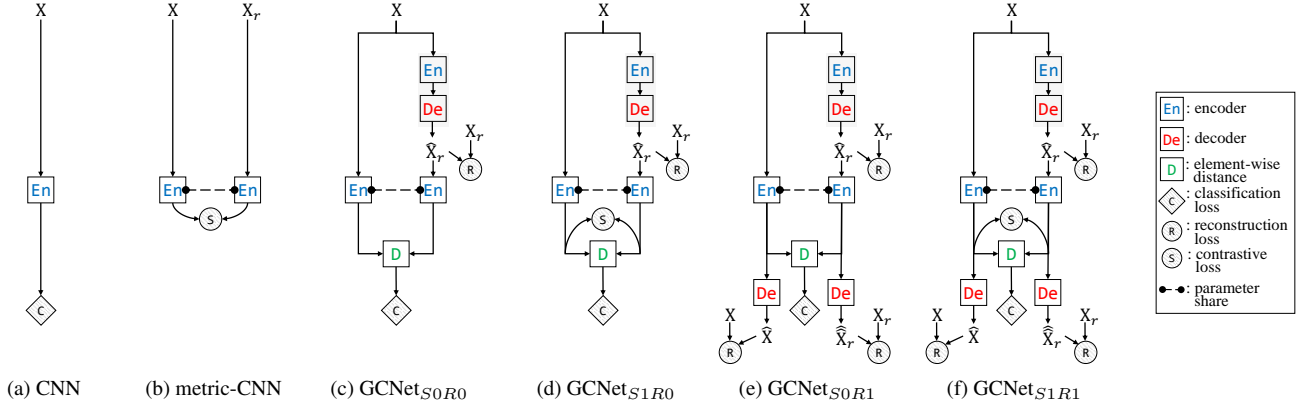


Figure 2: Architecture overviews of our proposed networks derived from (a) and (b) (in a training phase): (a) CNN, (b) metric-learning-CNN with a paired input $\{\mathbf{X}, \mathbf{X}_r\}$ where \mathbf{X} is a given expression and \mathbf{X}_r is a reference expression, (c) GCNet_{S0R0} : contrastive (distance) representation using a generative sample ($\hat{\mathbf{X}}_r$) for a discriminative task (where $\hat{\mathbf{X}}_r$ is a generated reference image via convolutional encoder-decoder networks), (d) GCNet_{S1R0} : a contrastive metric loss (S) is added on (c) GCNet_{S0R0} , (e) GCNet_{S0R1} : decoder networks with a reconstruction loss (R) are added on (c) GCNet_{S0R0} for a better representation ($\hat{\mathbf{X}}$ is a reconstructed sample of the given expression and $\hat{\mathbf{X}}_r$ is a reconstructed sample of the generated reference image), and (f) GCNet_{S1R1} : a contrastive metric loss (S) and a reconstruction loss (R) via decoder networks.

world, an expressive face might be changed from a reference ground face (due to emotional changes that incur facial muscle movements [25]). We define a relationship between two images with a hidden factor denoted by $\epsilon \in \mathbb{R}^{h \times w}$ formally as follows:

$$\mathbf{X} := \mathbf{X}_r + \epsilon \quad (1)$$

where the addition indicates operations for facial expression change¹.

As a facial expression is not always apparently represented as an absolute value, a quantity of expression change obtained by comparing with a reference image might be useful. An expression image with a very small change could be recognized via difference maps (e.g., a pixel-wise distance and optical flows). As a human keeps a neutral-like or less-expressive face most of time, that face image could be considered as the reference image.

A representation of a difference between expression images can appear in various ways. A simple approach is to compare image pixels of the faces. However, owing to distortions between the images (e.g., distortions by an affine transform), comparing the images at the pixel level is not effective. For example, a small translation in the image level might return large pixel-wise errors even though a human face shows no expression changes.

¹Since the change of expression should be measured in the same subject, we assumed that a hidden expression factor is represented within the same subject, i.e. if a subject term s is added at the Equation (1): $\mathbf{X}_s := \mathbf{X}_{s_r} + \epsilon$. In this paper, we omit the term s for a simplicity in the notation.

3.1. Representation of a difference between facial expressions via networks

The representation of the difference (“contrastive representation”) can be better extracted at the feature level, but not at the pixel level. The feature-wise representation can offer an invariance towards distortions (e.g., translation, scale, or rotation).

We employ a contrastive representation in the networks to extract a latent difference factor between expressions. Consider a pair of images $\{\mathbf{X}, \mathbf{X}_r\} \in \mathcal{I}$, where \mathbf{X} is an input sample and \mathbf{X}_r is a reference sample. Let En (abbreviation of an encoder) be a transform function used to map an input matrix to the embedding space ($\bullet \rightarrow \text{En}(\bullet) : \mathcal{I} \rightarrow \mathbb{R}^p$). In the transformed space, a latent factor in the feature level (δ) can be represented as follows:

$$\delta := d(\text{En}(\mathbf{X}), \text{En}(\mathbf{X}_r)), \quad (2)$$

where $d(\bullet, \bullet)$ is an element-wise distance formulation and $\delta \in \mathbb{R}^p$. In this paper, we adopt a distance $\|\text{En}(\mathbf{X})_j - \text{En}(\mathbf{X}_r)_j\| \in \mathbb{R}$ of the j -th element of the feature vector, ($\forall j = 1, \dots, p$) for $d(\bullet, \bullet)$.

In the contrastive representation of “expressiveness,” we expect that other factors (e.g., individual’s identity, pose, and etc.) than the expression will be eliminated. The contrastive representation δ is used for a discriminative task.

3.2. Generating reference image

The reference face image, such as less-expressive face image of an individual, may not be available in the test

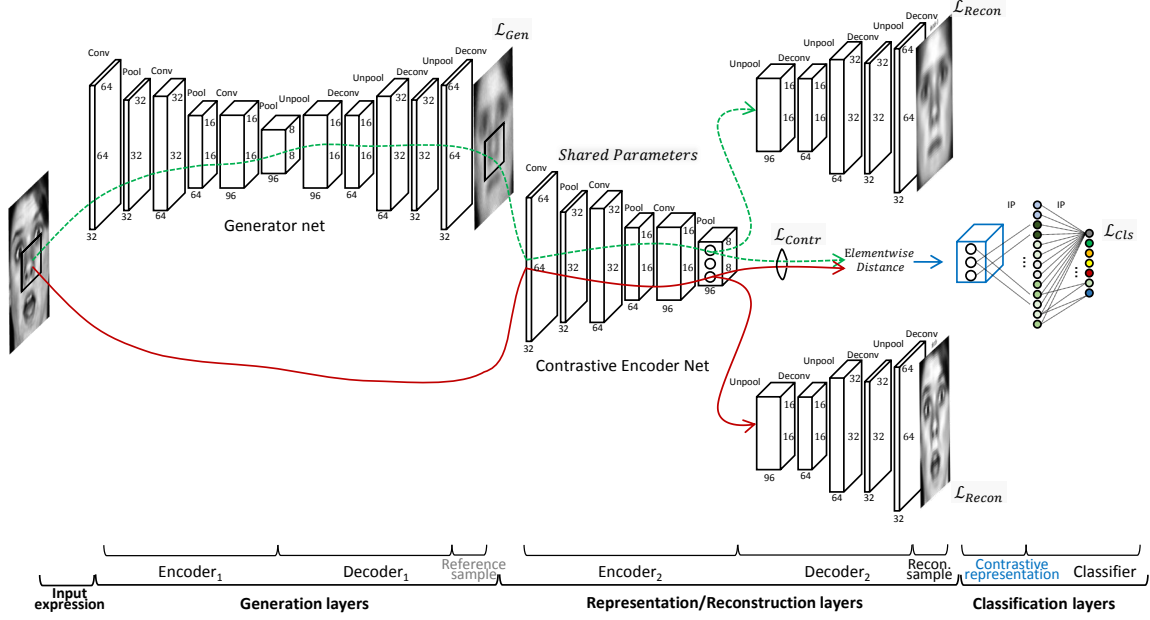


Figure 3: The overall architecture of the proposed networks (Figure 2 (f) in detail). Two-way data flows starting from a given expression image are existed over the generation layers and the representation/reconstruction layers. A dashed-line arrow (green color) depicts a flow to represent processing using a reference image generated by the generator networks. A solid-line arrow (red color) depicts a flow using the given expression image.

phase. In this paper, therefore, we propose to generate the reference face using convolutional encoder-decoder networks. To estimate a reference face transformed from an expressive face, we apply the concept of the denoising auto-encoder (DAE). In the DAE, a term corresponding to corruption, i.e., a Gaussian distribution, added to the original input is eliminated via learning² [3]. In this section, we assume that the term corresponding to corruption should not be limited to a specific probability distribution. There might be a latent model (or unknown transform) that makes a face with a certain expression appear to be a reference face. Without a definition of the latent distribution, in this paper, the model is represented using encoder-decoder networks. By disentangling facial expressive factors in feature learning, hence, information that is irrelevant or of negligible use for the discriminative purposes could be discarded [2].

3.3. Generative and contrastive networks

In this section, we show how the generated reference image can be used in deep networks. Multiple objectives are adopted to optimize parameters of the networks to generate a good contrastive representation.

As shown in Figure 2 (f) and 3, a loss function (\mathcal{L}) of

²An observed random variable X is corrupted into \tilde{X} using the known conditional distribution $C(\tilde{X} | X)$ in order to train the autoencoder to estimate the reverse conditional $P(X | \tilde{X})$

the proposed networks consists of three kinds of objectives. Formally, the loss function is written as follows:

$$\mathcal{L} = \mathcal{L}_{Cls} + \lambda_S \mathcal{L}_{Contr} + \lambda_R \mathcal{L}_{Recon} \quad (3)$$

where \mathcal{L}_{Cls} denotes a discriminative loss function, \mathcal{L}_{Contr} denotes a contrastive loss function, and \mathcal{L}_{Recon} denotes a reconstruction loss function. $\lambda_{\bullet} \in \mathbb{R}$ indicates a weight.

The main purpose of the proposed networks is to classify a facial expression in the given input. For the discriminative objective \mathcal{L}_{Cls} , we adopt the cross entropy loss function which is widely used for the classification task. Consider a pair of features $\{\text{En}_2(\mathbf{X}), \text{En}_2(\hat{\mathbf{X}}_r)\}$ extracted from encoder layers En_2 , where a subscript ₂ at En_2 indicates the second encoder layers (representation layers shown in Figure 3). A contrastive representation feature $d(\text{En}_2(\mathbf{X}), \text{En}_2(\hat{\mathbf{X}}_r)) \in \mathbb{R}^p$ where $d(\bullet, \bullet)$ is the element-wise distance and $p > 0$ is used for the classification task.

For learning a contrastive representation, two learning objectives are deployed in the proposed networks: the first objective is contrastive metric learning (\mathcal{L}_{Contr}) to enlarge or to diminish the distance between the two feature vectors, and the second is reconstruction learning (\mathcal{L}_{Recon}) for a better representation. Hence, the two objective functions are designed to jointly assist the classification task for realizing a good generalization performance.

Loss for contrastive metric learning in feature space.

The objective of the loss \mathcal{L}_{Contr} is to optimize a similarity between two features $\{\text{En}_2(\mathbf{X}), \text{En}_2(\hat{\mathbf{X}}_r)\}$ according to an expression label. If the expression labels of \mathbf{X} and $\hat{\mathbf{X}}_r$ are not identical, the function optimizes to obtain dissimilar features within a predefined margin; if the expressions are identical, it optimizes to similar features. Hence, the contrastive loss [10] is adopted for \mathcal{L}_{Contr} in a feature space as follows:

$$\mathcal{L}_{Contr} = \alpha \frac{1}{2} \{\max(0, m - S(\text{En}_2(\mathbf{X}), \text{En}_2(\hat{\mathbf{X}}_r)))\}^2 \quad (4)$$

$$+ (1 - \alpha) \frac{1}{2} \{S(\text{En}_2(\mathbf{X}), \text{En}_2(\hat{\mathbf{X}}_r))\}^2 \quad (5)$$

where $\alpha = 1$ if the labels of a pair $\{\mathbf{X}, \hat{\mathbf{X}}_r\}$ are not the same, $\alpha = 0$ otherwise, $S(\text{En}_2(\mathbf{X}), \text{En}_2(\hat{\mathbf{X}}_r)) = \|\text{En}_2(\mathbf{X}) - \text{En}_2(\hat{\mathbf{X}}_r)\|_2 \in \mathbb{R}$ is a similarity measure, and $m > 0$ is a margin. A feature space is defined as $(\bullet \rightarrow \text{En}_2(\bullet) : \mathcal{I} \rightarrow \mathbb{R}^p)$ at the encoder layers.

Loss for generation and representation. The main objectives of the loss \mathcal{L}_{Recon} are two-fold: one is to generate a reference image, and the other is to supplement to represent a good contrastive feature in the embedding layer ($\text{En}_2(\bullet)$). Hence, a reconstruction loss (\mathcal{L}_{Recon}) can be represented as a weighted summation of three terms as follows:

$$\mathcal{L}_{Recon} = \lambda_{G,r} \mathcal{L}_{Gen,r} + \lambda_{R,r} \mathcal{L}_{Recon,r} + \lambda_{R,i} \mathcal{L}_{Recon,i} \quad (6)$$

where of the subscripts of $\mathcal{L}_{*,\bullet}$, the first one $*$ $\in \{\text{Gen}, \text{Recon}\}$ indicates the stage for a generation (*Gen*) or a reconstruction (*Recon*) (shown in Figure 3). The second subscript $\bullet \in \{r, i\}$, indicates a target: r for a reference image, and i for an input image.

$$\mathcal{L}_{Gen,r} = \frac{1}{2} \|\mathbf{X}_r - \text{De}_{1,r}(\text{En}_1(\mathbf{X}))\|_2^2, \quad (7)$$

$$= \frac{1}{2} \|\mathbf{X}_r - \text{Generator}(\mathbf{X})\|_2^2, \quad (8)$$

$$\mathcal{L}_{Recon,r} = \frac{1}{2} \|\mathbf{X}_r - \text{De}_{2,r}(\text{En}_2(\text{Generator}(\mathbf{X})))\|_2^2, \quad (9)$$

$$\mathcal{L}_{Recon,i} = \frac{1}{2} \|\mathbf{X} - \text{De}_{2,i}(\text{En}_2(\mathbf{X}))\|_2^2, \quad (10)$$

where $\text{Generator}(\mathbf{X}) = \text{De}_{1,r}(\text{En}_1(\mathbf{X}))$ is learned to estimate \mathbf{X}_r .

4. Experiments

In this section, we describe the experiments conducted to compare the proposed method with the state-of-the-arts on three publicly available face expression databases (CK+, MMI, and Oulu-CASIA) that are widely adopted in the literatures [9, 12, 13, 14, 15, 16, 17, 18, 19, 23, 25, 26, 27, 29].

4.1. Networks model and settings

All models used in different databases share exactly the same architecture (shown in Figure 3), including encoder-decoder networks depicted in Table 1. All parameter settings are shared through the databases with the same value. The encoder-decoder networks in Table 1 are pre-trained with the reconstruction task using the CASIA-WebFace database [6], and three convolutional layers in the encoder are adopted at Encoder₁ (En_1) of the proposed generative-contrastive networks (GCNet) shown in the Figure 3. The baseline CNN consisting of three convolutional layers and two inner-product (FC) layers are pre-trained with the identification task using the same database, and convolutional layers are adopted at Encoder₂ (En_2). During the training of the proposed networks, the learning rate at layers of the decoder networks is set to 10 during fine-tuning. The number of outputs at the first fully-connected layer (inner-product) is empirically determined by $(0.5 * \text{Wsize})^2 * \text{nlayers} / (2^{\text{nlayers}})$ where we set $\text{Wsize} = 64$, $\text{nlayers} = 3$. This is intended that a dimensionality of the vector decreases smoothly as the number of (conv./pool) layers increases. $\frac{1}{2^{\text{nlayers}}}$ is related to a pooling size ($\frac{1}{2}$) at each layer. The dropout is applied before this fully-connected layer with a ratio of 0.5. After the FC-layer, a softmax layer is connected with the number of outputs corresponding to the number of classes. We arbitrarily set $\lambda_S = 1$, $\lambda_{G,r} = 1$, $\lambda_{R,r} = 0.25$, $\lambda_{R,i} = 0.25$ for each loss function. The maximum iteration is set to 3×10^5 .

Our models are trained with ‘Nesterov’ optimization using an ‘inverse’ learning policy, a base learning rate of 0.001, a momentum of 0.9, a gamma term of 0.75, a weight decay of 0.0001, and a mini-batch size of 64. The proposed network model is implemented on *Python* and the deep learning framework *Caffe* and run using the NVIDIA Tesla K80 GPU.

To avoid over-fitting, we applied data augmentation during the training phase. We used input images on a gray level (1 channel) where a facial region is cropped, normalized based on 5 points (eyes, the end of a nose, and two ends of lips) and resized into 66×66 . The resized image is cropped again with the size of 64×64 at a random location. Each cropped image is manipulated using 2D affine transform such as scaling, rotation, and intensity multiplication, in addition to random flipping.

4.2. Databases and protocols

CK+ Database [20] This database is widely adopted in the benchmark for facial expression recognition tasks. This database consists of 593 sequences with 123 individuals. The images are captured expression transitions from a neutral face to peak facial expression acted by an individual. The 327 valid sequences with 118 individuals that maintain discrete emotion labels such as ‘Anger, Contempt, Dis-

gust, Fear, Happy, Sad, and Surprise” are adopted for an experiment. We divide the valid sequences into ten different subsets with individual-independent way. According to individual ID in the database, individuals are grouped by sampling in ID ascending order with ten even intervals first. One subset out of ten subsets is used for validation (test), the remains are used for training. This procedure is repeated ten times. This 10-fold cross-validation follows the previous works [13, 19].

MMI Database [24] This database consists of 312 sequences from 30 individuals with six basic expressions (Contempt included in the CK+ database is excluded). We selected 205 sequences captured in a front view. Each sequence starts from a neutral face, and shows a peak expression within a single expression type in the middle of the sequence. At the end, it returns to a neutral face again. As a peak expression frame number is not given, we selected it manually. Similar to the CK+ database settings, we divided the MMI database into ten different individual independent subsets. Consequently, 10-fold cross validation was conducted. This database includes individuals who pose expressions non-uniformly, wear glasses/caps, and have mustaches/head movements. Therefore, the facial expression recognition task is relatively challenging. Moreover, the small number of sequences and individuals makes it difficult to achieve a good generalization performance. This database could be suitable to measure the recognition performance in realistic situations when compared to other databases.

Oulu-CASIA VIS Database [26] This database consists of 480 image sequences with 80 individuals. This database is captured under the visible (VIS) normal illumination conditions and is a subset of Oulu-CASIA NIR-VIS database. Each individual poses six basic expressions similar to MMI database. Similar to the CK+ database, the sequence starts from a neutral face and ends with peak facial expression within a same emotion category. As done with the two databases above, individual-independent 10-fold cross-validation is conducted.

4.3. Quantitative results

Among all the compared databases, the proposed methods outperform the state-of-the-art methods including handcraft based methods (LBP-TOP [27] and HOG 3D [15]), video-based methods (MSR [23], TMS [12], STM-ExpLet [19], and DTAGN-Joint [13]) that utilize temporal information, FAU inspired methods (AURF [16], AUDB [17]), and CNN-based methods (3D-CNN [19], 3D-CNN-DAP [19], zero-bias CNN+AD [14], and DTAGN-Joint [13]).

In the CK+ database, seven expressions and a neutral image are included. We conducted experiments for

Encoder (3 convolutional layers)
(5×5 , 32) Conv. BNorm, ReLU, (5×5) MaxPool
(3×3 , 64) Conv. BNorm, ReLU, (3×3) MaxPool
(3×3 , 96) Conv. BNorm, ReLU, (3×3) MaxPool
Decoder (3 de-convolutional layers)
(3×3) MaxUnPool, (3×3 , 32) DeConv. BNorm ReLU
(3×3) MaxUnPool, (3×3 , 64) DeConv. BNorm, ReLU
(5×5) MaxUnPool, (5×5 , 1) DeConv. BNorm, ReLU

Table 1: Details of the convolutional encoder-decoder layers [22] embedded in the proposed networks. An encoder part consists of three convolutional layers (Conv.) which is followed by Batch Normalization (BNorm), ReLU, and Max Pooling layers. Correspondingly, a decoder part consists of three de-convolutional (transposed convolutional) layers. In a Conv and DeConv. layers, (5×5 , 32) indicates that there is 32 sets of 5×5 filters. In MaxPool and MaxUnPool layers, (5×5) indicates a pooling window size.

seven expressions as well as eight expressions (seven expressions and a neutral face). For the seven expressions cases shown in Table 2, the proposed methods (GCNet_{S0R0}, GCNet_{S1R0}, GCNet_{S0R1}, and GCNet_{S1R1}) show a better recognition performance than that of all compared state-of-the-arts including hand-craft feature based methods (LBP-TOP [27] and HOG 3D [15]), CNN-based methods (3D-CNN [19], 3D-CNN-DAP [19], and DTAGN-Joint [13]), and video-based methods (MSR [23], TMS [12], STM-ExpLet [19], and DTAGN-Joint [13]). For cases of the eight expressions shown in Table 3, the proposed methods (GCNet_{S0R0}, GCNet_{S1R0}, GCNet_{S0R1}, and GCNet_{S1R1}) show a better recognition performance than the compared deep learning-based methods including FAU aware methods (AURF [16], AUDB [17]) and a CNN-based method (Zero-bias CNN+AD [14]). When a loss function of contrastive metric learning is eliminated (GCNet_{S0R0} and GCNet_{S0R1}), we observed that the performance is degraded than that with a contrastive loss (GCNet_{S1R0} and GCNet_{S1R1}) on the CK+ database.

In the MMI database, similar to the case of the CK+ database, the proposed methods show a higher accuracy value than that of the state-of-the-arts including CNN-based methods (3D-CNN-DAP [19] and DTAGN-Joint [13]) and video-based methods (STM-ExpLet [19] and DTAGN-Joint [13]) as shown in Table 4. The methods (STM-ExpLet [19] and DTAGN-Joint [13]) that acquire temporal information from multiple images show relatively higher accuracy performance than other methods. Even though the proposed methods show a better recognition performance than these compared methods, the recognition accuracy of the proposed methods on the MMI database is relatively less compared to that on other databases (CK+ and Oulu-CASIA

Method	Accuracy (%)
LBP-TOP [27]	88.99
HOG 3D [15]	91.44
MSR [23]	91.4
TMS (4-fold) [12]	91.89
STM-ExpLet [19]	94.19
DTAGN-Joint [13]	97.25
3D-CNN [18]	85.9
3D-CNN-DAP [18]	92.4
CNN (baseline)	96.94
Ours (GCNet _{S0R0})	97.08
Ours (GCNet _{S1R0})	97.83
Ours (GCNet _{S0R1})	97.53
Ours (GCNet _{S1R1})	97.93

Table 2: Averaged recognition accuracy (%) on the CK+ database, 7 expressions.

Method	Accuracy (%)
AURF [16]	92.22
AUDB [17]	93.70
Zero-bias CNN+AD [14]	96.4
CNN (baseline)	95.47
Ours (GCNet _{S0R0})	95.74
Ours (GCNet _{S1R0})	96.75
Ours (GCNet _{S0R1})	96.50
Ours (GCNet _{S0R0})	97.28

Table 3: Averaged recognition accuracy (%) on the CK+ database, 8 expressions.

VIS). Due to the large intra-identity variation of the MMI database, locally selected patch based method (CSPL [29],) shows a relatively better performance than other compared methods.

In the Oulu-CASIA VIS database, similar to the CK+ case, the proposed methods show a higher accuracy value than the state-of-the-arts including CNN-based methods (DTAGN-Joint [13]) and video-based methods (AdaLBP [26], Atlases [9], STM-ExpLet [19] and DTAGN-Joint [13]), as shown in Table 5.

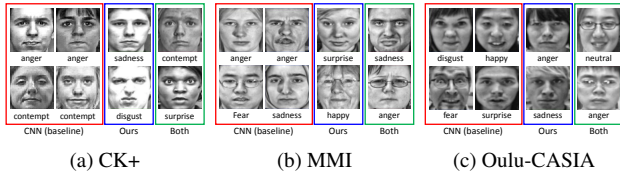


Figure 4: Recognition errors on small depth of expression cases.

Method	Accuracy (%)
LBP-TOP [27]	59.51
HOG 3D [15]	60.89
ITBN [25]	59.7
CSPL [29]	73.53
STM-ExpLet [19]	75.12
DTAGN-Joint [13]	70.24
3D-CNN [18]	53.2
3D-CNN-DAP [18]	63.4
CNN (baseline)	77.68
Ours (GCNet _{S0R0})	76.20
Ours (GCNet _{S1R0})	78.86
Ours (GCNet _{S0R1})	77.00
Ours (GCNet _{S1R1})	81.53

Table 4: Averaged recognition accuracy (%) on the MMI database, 6 expressions.

Method	Accuracy (%)
LBP-TOP [27]	68.13
HOG 3D [15]	70.63
AdaLBP [26]	73.54
Atlases [9]	75.52
STM-ExpLet [19]	74.59
DTAGN-Joint [13]	81.46
CNN (baseline)	83.96
Ours (GCNet _{S0R0})	84.65
Ours (GCNet _{S1R0})	86.39
Ours (GCNet _{S0R1})	85.83
Ours (GCNet _{S1R1})	86.11

Table 5: Averaged recognition accuracy (%) on the Oulu-CASIA VIS database, 6 expressions.

4.4. Qualitative analysis

Small expression images In Figure 4, several examples of recognition errors on the small expression images are shown. Our proposed method shows approximately twice less recognition errors than baseline CNN method. However, both the baseline CNN and the proposed method still have a limitation to recognize the small (or ambiguous) expressions.

Visualization in the feature space To observe a discriminative distribution of the extracted features, we visualized the feature vectors from the first layer of the fully-connected layers of the proposed networks of the baseline CNN and our proposed networks. We visualize the 384 dimensional feature vectors using t-SNE [21]. The feature points of original images are scattered within a narrow region. The point distribution of the baseline CNN forms partially overlapped

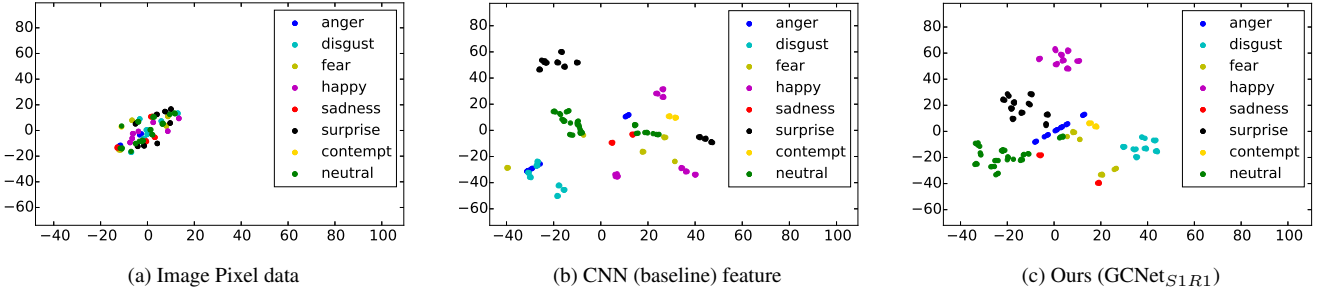


Figure 5: Visualization of the extracted features using t-SNE: (a) a pixel value of the input images, (b) a feature vector of CNN (baseline), and (c) a feature vector of the proposed method (GCNet_{S1R1}).

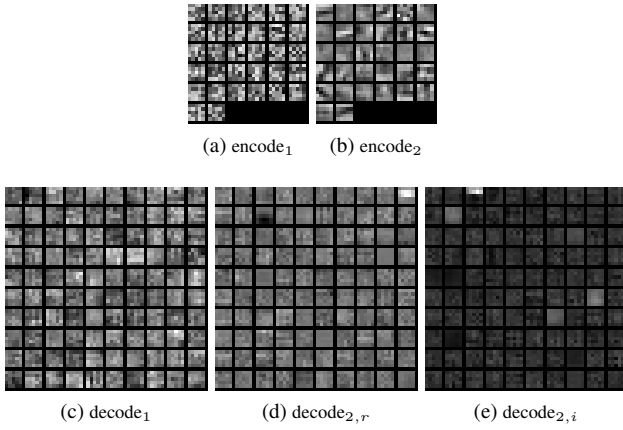


Figure 6: Learned filter examples on the CK+ database. (a) Conv. filters in En₁, (b) Conv. filters in En₂, (c) DeConv. filters in De₁, (d) DeConv. filters in De_{2,r}, and (e) DeConv. filters in De_{2,i}: (a) and (b) are from the first Conv. layer’s filters (32 filters) at the encoder networks, and (c), (d), and (e) are from the second DeConv. layer (100 filters randomly selected) at the decoder networks.

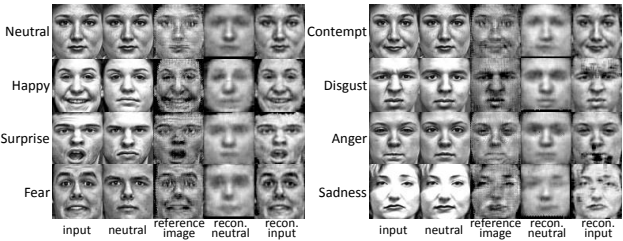


Figure 7: Examples of generation and reconstruction results on the test data.

clusters. The proposed network features are clustered well to discriminate individual expression further.

Patterns of the learned filters We observe the characteristics of the filters learned in the proposed networks. As shown in Figure 6, the encode filters learned by contrastive metric learning, (b), has more Gabor like edge and blob detection filters than (a). The decoder filters for the expression reconstruction, (e), show a simpler patterns than that for the reference image generator, (c), and the reconstruction decoder of a neutral image, (d), as shown in Figure 6.

Visualization of the response maps We observe the response maps resulted from generation and reconstruction layers of the proposed networks to understand what the networks have been conducted in the test phase. In Figure 7, a generated reference image, a reconstructed neutral image, and a reconstructed image of a given expression are shown. The generated reference image is affected by reconstruction and contrastive metric learning.

5. Conclusions

In this paper, we proposed facial expression recognition method based on contrastive representation learning. The contrastive representation is calculated in the embedding layer of deep networks by comparing a single given image with a reference image. The reference image is generated by deep generative (encoder-decoder) networks. This approach is useful especially if an expressive depth of an emotional face is varied among individuals, expressions, or situations. In our proposed networks, we attempted to disentangle a facial expressive factor directly. Disentangling of expression is conducted in two steps: 1) learning of a reference face by a generator network and 2) learning of a contrastive representation with a combination of contrastive and reconstruction objectives. Extensive experiments were conducted on three face expression databases that are publicly available and widely adopted in the literature. The proposed method outperforms the known state-of-the arts, including both single image and multiple-image based meth-

ods. This study could be extended to effectively detect and recognize small changes of facial expressions from sequential images.

References

- [1] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573. IEEE, 2005. 2
- [2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013. 4
- [3] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13*, pages 899–907, 2013. 4
- [4] V. Bruce and A. Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986. 1
- [5] A. J. Calder and A. W. Young. Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8):641–651, 2005. 1
- [6] S. L. Dong Yi, Zhen Lei and S. Z. Li. Learning face representation from scratch. In *arXiv preprint arXiv:1411.7923*. 2014. 5
- [7] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978. 2
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014. 2
- [9] Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *Computer Vision–ECCV 2012*, pages 631–644. Springer, 2012. 5, 7
- [10] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 5
- [11] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, 2000. 1
- [12] S. Jain, C. Hu, and J. K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1642–1649. IEEE, 2011. 5, 6, 7
- [13] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015. 2, 5, 6, 7
- [14] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015. 2, 5, 6, 7
- [15] A. Klser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *In BMVC08*, 2008. 2, 5, 6, 7
- [16] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013. 2, 5, 6, 7
- [17] M. Liu, S. Li, S. Shan, and X. Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159:126–136, 2015. 2, 5, 6, 7
- [18] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*, pages 143–157. Springer, 2014. 2, 5, 7
- [19] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. 2, 5, 6, 7
- [20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010. 5
- [21] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 7
- [22] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 6
- [23] R. W. Ptucha, G. Tsagkatakis, and A. E. Savakis. Manifold based sparse representation for robust expression recognition without neutral subtraction. In *ICCV Workshops*, pages 2136–2143. IEEE, 2011. 5, 6, 7
- [24] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010. 6
- [25] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013. 3, 5, 7
- [26] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 5, 6, 7
- [27] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, June 2007. 2, 5, 6, 7
- [28] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked What-Where Auto-encoders. *arXiv*, 2015. 2

- [29] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2562–2569. IEEE, 2012.
- 5, 7